# Multi-class Imbalanced Learning

DingHao

December 5, 2016

# Contents

Introduction

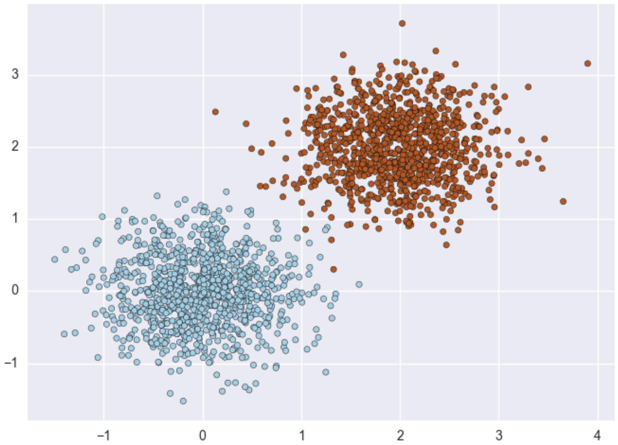
Evaluation Criteria

Approaches

Future Work

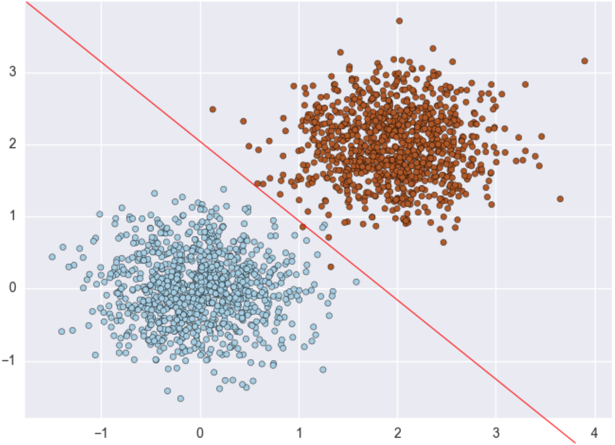
# Introduction

## Classification Problem



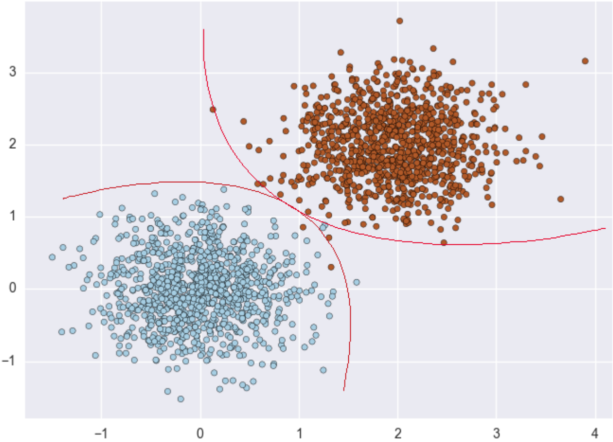
# Introduction

## classifiers



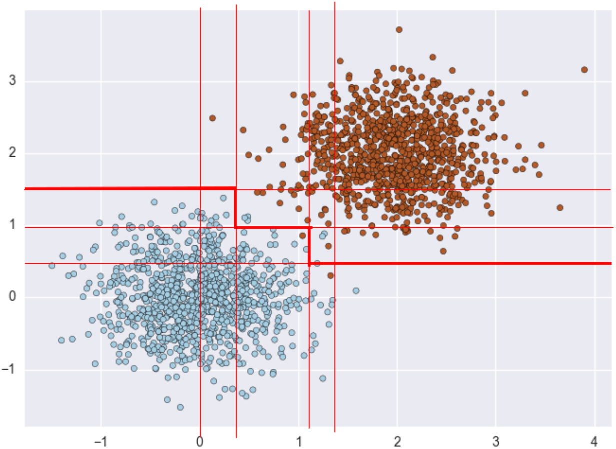
# Introduction

## classifiers



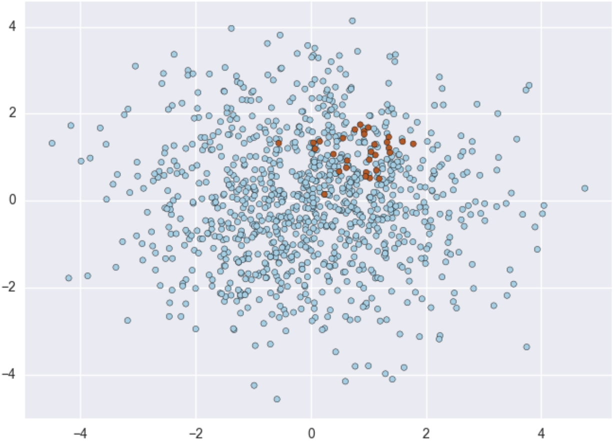
# Introduction

## classifiers



# Introduction

## Real World

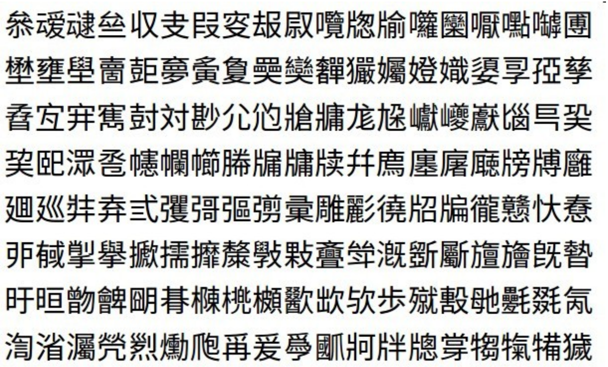


# Introduction

## Doggy

# Introduction

## Rare Word



叅𥕦𥖽坐収攴叚安赧叞嚂牎牏囉圞嚈嚸嘑團

壄壅壆亶距夢禽复哭奱鞸獵孎嬁嬂娎享掗孿

昚宜宑寯尌对尠㕣伀牄牗龙尵巘巙巚𡿺甹契

契巸罛吞㦜㦨𢣐脪牖牅牍幷鴈廛屠廳牓牔匯

廻廵丰奔弎彏彁彄𢏗彚雕彲徺招牑徿戇忕惷

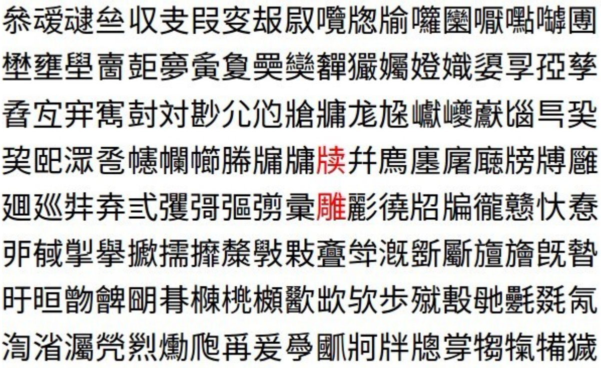
夘椷𦥯擧擨擩擵蓁斆敤亹丵溉斱斸旜旝旣朁

旴晅𪟝鞞朙朞楝桄櫴歠欪欤歩殧殼毑氎毲氝

淘消灟党烈爋爮爯爰爭鄜牁牉牕掌犓𤛓犕獩

# Introduction

## Rare Word



叅𥔲逮坐収叏叚安赧叞𡅏牕牏囉圝嚈嚸嚹團
壄埀壆畗䝳夢黃复噐欒鞸獵孎嬁嬂娶享硜攣
昋亙宑雋封对尠允㳂牄牅龙尳巘巙巚𡿺号契
契巸𣶒昬㥀幱幯幐牑牗牍并廌廛屠廳膀膊廲
廻廵弉奔弎彏彁彄彅彙雕彲徺卲牑徿竸忕惷
戼槭𢸜擧攇擩攠檕斅敕亹竿溉斲斸旜旝旣朁
旰晅𩐁錍朙朞楝桃櫴歠欪欤步殧殼毑氈毲氝
淘消灟党𤋮爋爮爯爰孚𤖱牁牉牕牚犓𤚩犕獩

# Introduction

## Imbalanced Ratio

imbalanced ratio = majority class / minority class

### ZooScan

427 / 28 = 15.25

### Kaggle

1979 / 9 = 219.89

### WHOI

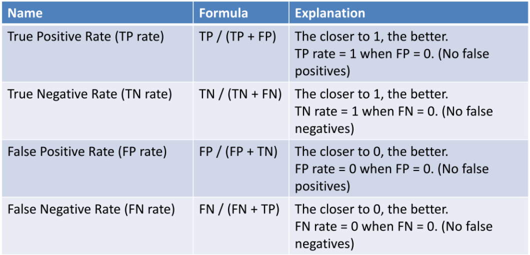
2606720 / 4 = 651680

EVEN MORE THAN $10^8$ !

# Introduction

WHY?

# Evaluation Criteria



| Name | Formula | Explanation |
|---|---|---|
| True Positive Rate (TP rate) | TP / (TP + FP) | The closer to 1, the better.<br>TP rate = 1 when FP = 0. (No false positives) |
| True Negative Rate (TN rate) | TN / (TN + FN) | The closer to 1, the better.<br>TN rate = 1 when FN = 0. (No false negatives) |
| False Positive Rate (FP rate) | FP / (FP + TN) | The closer to 0, the better.<br>FP rate = 0 when FP = 0. (No false positives) |
| False Negative Rate (FN rate) | FN / (FN + TP) | The closer to 0, the better.<br>FN rate = 0 when FN = 0. (No false negatives) |

$$G - mean = \sqrt{TPr * TNr}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN} = TPr$$

$$F - measure = \frac{2 * Precision * Recall}{Precision + Recall}$$

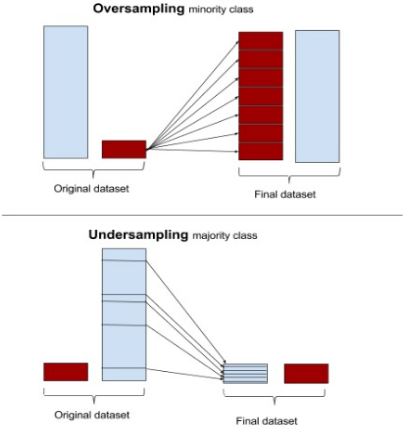
# Approaches

## Overview

- Sampling
  *Under-sampling*
  *Over-samping*
- Cost-sensitive learning
- Ensembled classifier
  *EasyEnsemble*
  *BalanceCascade*

# Approaches

## Sampling


**Oversampling** minority class
Original dataset
Final dataset
**Undersampling** majority class
Original dataset
Final dataset

Best approache: SMOTE

# Approaches

## Cost-sensitive

$$L(x, i) = \sum_{j} P(j|x) c(i, j)$$

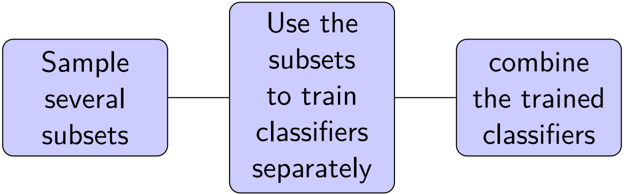
Minimize the overall cost.

- x : an example
- i : a class
- j : the j$^{th}$ class
- P : Probability
- c : cost matrix

Best approaches: AdaCost, AsymBoost

# Approaches

## Ensembled Classifier


Sample several subsets

Use the subsets to train classifiers separately

combine the trained classifiers

Best approaches: EasyEnsemble, BalanceCascade, SMOTEBoost

# Approaches

## EasyEnsemble.M ⇒ EasyEnsemble.D

```
Input:A set of minority class examples $\mathcal{P}$, k-1 sets of majority class examples $\mathcal{N}$, $|\mathcal{P}| < |\mathcal{N}_k|$, the number of subsets T to sample from $\mathcal{N}_k$, and $s_i$,the number of iterations to train an AdaBoost ensemble $H_i$
for i⇐ 1:T
    $D_i = \mathcal{N}_1$
    for t⇐ 1:k
        Randomly sample a subset $\mathcal{N}_{it}$ from $\mathcal{N}_k$, $N_{it}$,$|N_{it}| = |P| + \frac{\mathcal{N}_1 * (|\mathcal{N}_i| - |P|)}{|\mathcal{N}_k|}$ in the $t^{th}$
        $D_i = D_i \bigcup \mathcal{N}_{it}$
    $H_t\left(x\right) = sgn\left(\sum_{d=1}^{s_i} \alpha_{t,d} h_{t,d}\left(x\right) - \theta_i\right)$
$H\left(x\right) = sgn\left(\sum_{t=1}^{T} \sum_{d=1}^{s_i} \alpha_{t,d} h_{t,d}\left(x\right) - \sum_{t=1}^{T} \theta_i\right)$
```

Q.-Q.Li, and X.-Y.Liu. "EasyEnsemble.M for multiclass imbalance problem." In: Pattern Recognition and Artificial Inte
27.2(2014):187-192.

## Future Work

- Optimize the algorithm to cost less runtime
- Use Kaggle and WHOI datasets
- Increase the amount of time in each dataset

Q&A